(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
21 March 2002 (21.03.2002)

PCT

(10) International Publication Number
WO 02/23814 A2

(51) International Patent Classification[7]: H04L 12/00

(21) International Application Number: PCT/US01/42072

(22) International Filing Date:
7 September 2001 (07.09.2001)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
09/659,106 11 September 2000 (11.09.2000) US

(71) Applicant: SUN MICROSYSTEMS, INC. [US/US]; 901 San Antonio Road, MS UPAL01-521, Palo Alto, CA 94303 (US).

(72) Inventors: EBERLE, Hans; 464 Dell Avenue, Mountain View, CA 94043 (US). GURA, Nils; 450 Oak Grove Dr. #305, Santa Clara, CA 95054 (US).

(74) Agents: ZAGORIN, Mark et al.; Zagorin, O'Brien & Graham, LLP, Suite 870, 401 West 15th Street, Austin, TX 78701 (US).

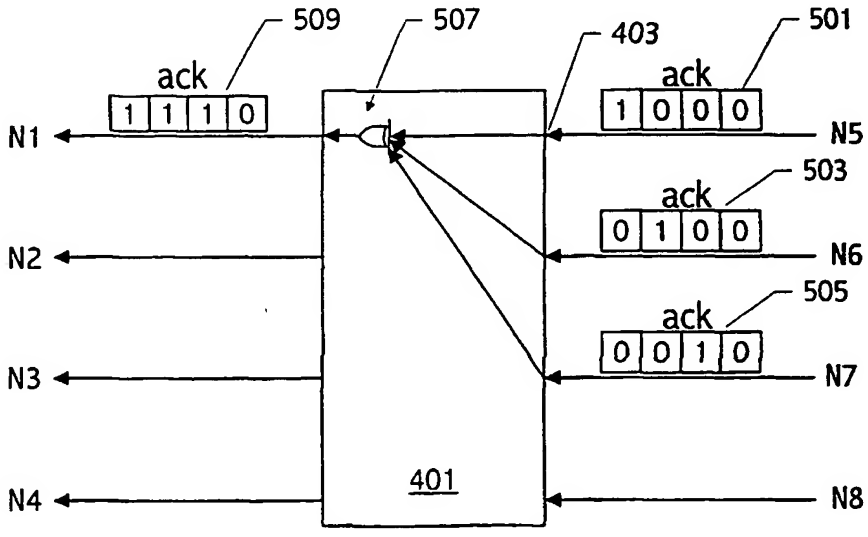
(81) Designated States *(national)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, PH, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, TZ, UA, UG, UZ, VN, YU, ZA, ZW.

(84) Designated States *(regional)*: ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE, TR), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:
— *without international search report and to be republished upon receipt of that report*

[Continued on next page]

(54) Title: RELIABLE MULTICAST USING MERGED ACKNOWLEDGMENTS



(57) Abstract: A source multicasts information to a plurality of targets. The targets respond to the multicast information by sending acknowledgments that indicate receipt of the multicast information. The acknowledgments are merged into a merged acknowledgment, which is then supplied to the source. The source can determine from the merged acknowledgment whether the targets successfully received the multicast information.

WO 02/23814 A2

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

RELIABLE MULTICAST USING MERGED ACKNOWLEDGMENTS

Technical Field

The invention relates to communication of information and more particularly to multicast operations.

Background Art

In current computing environments, especially networked environments, a source node on the network may wish to supply a plurality of destination nodes with the same information. In such situations, some systems provide a multicast capability in which the source node can send multiple destination nodes the same information at the same time. In such multicast operations, any number of multiple targets can receive the multicast information.

Referring to Fig. 1, a multicast operation is illustrated in which an initiator node I_0 simultaneously sends the same information to target nodes T_0, T_1, and T_2. Because the destination or target nodes can receive the multicast information simultaneously, the multicast operation is time efficient.

One difficulty with multicasting simultaneous information is that it may be difficult for the initiator node who sends the information to determine if the target nodes successfully received the information. Thus, the operation is unreliable in the sense that the initiators cannot determine if the transmission was successful. If the receiving nodes send acknowledgments indicating successful receipt of the multicast information, there would be a tendency for the acknowledgments to collide or otherwise contend for resources of the communication medium. That is because the targets would likely send the acknowledgments to the initiator node at the same time. In a switched synchronous network, sending such acknowledgments could result in undesirable collisions and possible loss of acknowledgment information. In other systems, the acknowledgments may be buffered within the switch as collisions occur, or require retry as some targets would be unable to obtain the communication medium to send the acknowledgment. In either of those situations, the advantage of time efficiency is diminished if acknowledgments take a long time relative to the original multicast due to contention for resources of the communication medium connecting the sending and receiving nodes.

One way to avoid such contentions and/or collisions is to provide the information sequentially as shown in Fig. 2, rather than simultaneously, as shown in Fig. 1. In the sequential operation, the initiator node I_0 successively sends the same information at 201, 202 and 203 to the target nodes T_0, T_1, and T_2. The target nodes respond sequentially with acknowledgments at 204, 205 and 206. Because the acknowledgments are sequential, they do not compete with each other for communication medium resources. Thus, the operation is reliable in the sense that the initiator can determine if the transmission was successful. However, the sequential nature of the operation for both the transmission of the information and the transmission of the acknowledgments eliminates any efficiency which could be gained from a true multicast operation in which multicast information is sent simultaneously. Thus, there is a relatively long latency for completion of the entire operation.

For certain time-critical multicast operations, it is important to minimize latency. For example, for time-critical multicast operations such as synchronization of clocks in a network, coherency protocols, and operations in databases/transaction systems such as *commit* or *abort*, minimizing latency would be advantageous.

Accordingly, it would be desirable to provide a multicast operation that is both efficient and reliable.

DISCLOSURE OF THE INVENTION

Accordingly, in one embodiment, the invention provides a method of multicasting that simultaneously sends multicast information from a source to a plurality of targets. The targets respond to the multicast information by sending acknowledgments that indicate receipt of the multicast information. The acknowledgments are merged into a merged acknowledgment, which is then supplied to the source. The source can determine from the merged acknowledgment whether the targets successfully received the multicast information.

In an embodiment, the multicast information and acknowledgments are transmitted across a network switch and the switch merges the acknowledgments before forwarding the merged acknowledgment to the source.

In another embodiment, a method is provided for transmitting information between an initiator node in a network and a plurality of target nodes. The method includes transmitting information from the initiator node to the target nodes simultaneously; simultaneously sending acknowledgments from the multiple nodes indicating receipt of the information; combining the acknowledgments and sending the combined acknowledgments to the initiator node to indicate receipt of the multicast information by the target nodes.

In another embodiment, the invention provides a data network that includes a sending node and a plurality of receiving nodes coupled to simultaneously receive information from the sending node during a multicast operation and coupled to respectfully provide acknowledgments of successful receipt of the multicast information. A switching medium supplies the multicast information to the respective receiving nodes simultaneously. Logic in the switching medium receives and combines the respective acknowledgments to provide a combined acknowledgment to the sending node. The combined acknowledgment may be a logical combination of the individual acknowledgments.

BRIEF DESCRIPTION OF THE DRAWINGS

The present invention may be better understood, and its numerous objects, features, and advantages made apparent to those skilled in the art by referencing the accompanying drawings.

Fig. 1 illustrates operation of an unreliable multicast operation in which no acknowledgments are provided by the targets.

Fig. 2 illustrates a sequential operation.

Fig. 3 illustrates operation of a reliable simultaneous multicast operation.

Fig. 4 illustrates an embodiment in which a multi-port switch is used for a multicast operation.

Fig. 5 illustrates an embodiment in which a multi-port switch is used to merge the acknowledgments, to indicate successful completion of the multicast operation.

Fig. 6 illustrates an embodiment in which a multi-port switch is used to merge the acknowledgments, to indicate a failed multicast operation.

Fig. 7 illustrates how single-bits can be concatenated into a vector.

Fig. 8 illustrates the transmission portion of a merged single-bit acknowledgment approach.

Fig. 9 illustrates the acknowledgment portion of the merged single-bit acknowledge operation.

The use of the same reference symbols in different drawings indicates similar or identical items.

DESCRIPTION OF THE PREFERRED EMBODIMENT(S)

Referring to Fig. 3, operation of a reliable multicast operation is illustrated. Assume the system includes multiple nodes including the illustrated initiator node I_0 and three target nodes T_0, T_1 and T_2. The initiator node I_0 sends information (data) to the three targets T_0, T_1 and T_2 simultaneously, i.e., the initiator node I_0 multicasts the information to the three targets. Each target, assuming successful receipt, sends back an acknowledgment (ack) to the initiator node I_0. As described further herein, in order for the initiator node I_0 to receive the simultaneously sent acknowledgments, the acknowledgments are merged and then provided to the initiator node. The merger operation is described further herein.

Referring to Fig. 4, the first part of a reliable multicast operation according to an embodiment of the invention is illustrated. In the first part of the multicast operation, the multicast information in the form of packet(s) P, is sent from initiator node N1 through input port 403 to target nodes N5, N6 and N7 across multiport switch 401. Note that packet(s) P may be one or more packets comprising one or more bytes of data and/or control information.

Referring to Fig. 5, the acknowledge phase of the multicast operation is illustrated. Nodes N5, N6 and N7, which received the multicast packet(s) P, respectively send acknowledge packets (ack) 501, 503 and 505 to node N1, which sent the multicast packet(s) P. Note that the exemplary acknowledge packets are shown in simplified form without information such as address, type of operation or other control information that would typically be associated with such a packet. Further note that a host typically contains both an initiator node and a target node and that the initiator and target share the input and output port of the switch. For example, N1 and N5 belong to the same host and send packets to input port 403 and receive packets from output port 405.

The exemplary multiport switch 401 includes four possible inputs and four possible outputs. Thus, in the embodiment illustrated in Fig. 5, the acknowledge packet (ack) from each multicast target node includes a vector of four bits, one bit corresponding to one of four possible output ports or targets on the switch. As illustrated in Fig. 5, the leftmost bit in the vector corresponds to node N5, the next bit to node N6, etc. Thus, when node N5 acknowledges the multicast, it sets the leftmost bit in its acknowledge vector 501 to indicate that N5 successfully received the multicast packet(s) P. Node N6 sets the bit second from the left in its acknowledge vector 503 to indicate that it successfully received the multicast packet(s) P. Node N7 sets the bit third from the left in its acknowledge vector 505.

Output port 507 merges the acknowledge packets received respectively from nodes N4, N5 and N6. As illustrated in Fig. 5, that can be accomplished by ORing together the acknowledge packets in OR logic in output port 507. When ORed together the merged acknowledgment packet 509 is generated and supplied to node N1. Node N1 can determine from the three bits set in merged acknowledge packet 509 that nodes N5, N6 and N7 successfully received the multicast packet(s) P. Thus, multiport switch 401 can provide a reliable and efficient multicast operation, since the acknowledge packets can be sent over the switch efficiently. That is made possible by the merging implemented in the output port.

Referring to Fig. 6, another operation of the multicast acknowledge is illustrated when some of the target nodes of the multicast operation fail to correctly receive the multicast packet P. That may be the result of, e.g., uncorrectable errors detected by the receiving node. As can be seen, only node N6 correctly received the multicast packet(s) P as indicated by the "0100" in its acknowledge packet. When the acknowledge packets from N5, N6 and N7 are ORed together, merged acknowledge packet 601 results which indicates that errors were detected by two nodes (N5 and N7). Using that information, the node initiating the multicast node can take appropriate action in response to the detected errors, such as resending the multicast packet P to the nodes that failed.

As would be known in the art, there are many other ways to encode the sources of the acknowledgments and to merge the acknowledge packets. For example, while the OR operation is possible, an embodiment could simply select the relevant bit from each output port acknowledge vector for inclusion in a merged acknowledge vector. Referring to Fig. 7, an example is shown in which single bits from each of the targets is merged into a vector. More particularly, each bit 701, 702 and 703 is concatenated to form vector 704, which is presented to the source to indicate which targets successfully received the multicast data. Alternatively, the switch could provide a count of the number of acknowledging multicast targets that indicated successful receipt, although that implementation would likely require more logic.

In a typical system, the input ports (or the control logic associated with the input ports) are aware of the multicast operation from information contained in a packet header. From that information, the control logic knows to connect the input port to the appropriate output ports. There are various approaches that could be used to alert the output port to merge the acknowledgments received by the input ports from the various targets. For example, an acknowledge packet may be marked as a multicast acknowledgment. Assuming that the packets to be merged arrive at the input ports simultaneously, the output port merges those packets that are

destined for it and appropriately marked. Alternatively, e.g., in a pipelined network, the switch can remember that it scheduled a multicast data transfer and merge the acknowledge packets at a particular pipeline stage in the future. It is also possible for acknowledge packets destined for the same port to merge packets whenever there exists multiple acknowledge packets for the same output port. That assumes that acknowledge packets to be merged arrive simultaneously. Thus, a multicast acknowledge would be presumed in such situations. Note that the switch settings for forwarding the acknowledgments can be inferred from settings for forwarding the multicast data.

It is also possible to merge acknowledge packets into an acknowledge packet containing a single bit rather than a bit vector, which is then forwarded on to the initiator node. Atomic operations are one application for a merged single bit acknowledge. Referring to Figures 8 and 9, operation of a merged single bit acknowledge is illustrated. In Figure 8 a multicast operation sends data from initiator node N1 to target nodes N5, N6 and N7. A forwarding mask 801 is generated that indicates which of the possible targets received the multicast data. That forwarding mask is utilized in merging the acknowledgments into a single bit as illustrated in Figure 9.

Referring to Figure 9, node N5 sends back acknowledgment 901, node N6 sends back acknowledgment 902, and node N7 sends back acknowledgment 903 as shown. Note that acknowledgment 902 indicates that node N6 failed to properly receive the multicast data. The merging is accomplished as follows. The individual acknowledgments are inverted and logically combined in AND gates 904 with the forwarding mask 801. The output of AND gates 904 are then logically combined in NOR gate 905 to provide the single bit acknowledgment 906 to the initiating node N1. In the example illustrated in Figure 9, the zero acknowledgment 902 from node N6 causes the single bit acknowledgment to be a zero indicating that a failure occurred. Note that while the acknowledgments 901, 902, and 903 from nodes N5, N6 and N7 are shown as single bits, as one of ordinary skill in the art would understand the acknowledgments can be in the various forms, e.g., an acknowledge packet indicating successful receipt or an acknowledge packet indicating unsuccessful receipt (NACK). Further, the acknowledgment 906 can also be in the form of an acknowledge packet indicating successful transmission or no acknowledge (nack) packet indicating transmission failure. An important aspect of this embodiment is that the overall success or failure of the multicast is encoded in a single bit (or bits) without providing information regarding individual multicast success or failure of the targets.

Other acknowledgment variations are also possible. For example, fine-grained acknowledgments may be used in which separate bits are provided, e.g., for CRC error, permission error, buffer overflow, etc. Thus, an exemplary system combines the individual bits, e.g., for CRC error, for all the acknowledging targets. Again, individual bits can be merged into either a bit vector or a single bit. In the later case, one bit of the merged acknowledgments represent the CRC errors from all the targets, one bit represents all the permission errors etc. The initiator node would know whether or not all targets successfully received the packet with or without a CRC error, or permission error, etc.

Thus, an efficient and reliable multicast operation has been described. While described in relation to a multiport switch, any switching medium that can effectively merge the multicast acknowledges can effectively utilize the invention described herein.

The embodiments described above are presented as examples and are subject to other variations in structure and implementation within the capabilities of one reasonably skilled in the art. The details provided above should be interpreted as illustrative and not as limiting. Variations and modifications of the embodiments disclosed herein, may be made based on the description set forth herein, without departing from the scope and spirit of the invention as set forth in the following claims.

WHAT IS CLAIMED IS:

1. A method of multicasting, comprising:
sending multicast information from a source to a plurality of targets;
sending respective acknowledgments from each of the targets, indicating receipt of the multicast information;
merging the respective acknowledgments into a merged acknowledgment; and
supplying the merged acknowledgment to the source.

2. The method as recited in claim 1 wherein the multicast information is sent across a switch to a plurality of targets.

3. The method as recited in claim 2 wherein the respective acknowledgments are sent from the respective targets to the switch.

4. The method as recited in claim 3 wherein the switch merges the respective acknowledgments and forwards the merged acknowledgment to the source.

5. The method as recited in claim 4 wherein the acknowledgments are supplied in an acknowledgment packet encoding an identity of the acknowledging target.

6. The method as recited in claim 3 wherein the switch is a synchronous switch and all acknowledgments are received by the switch at the same time.

7. The method as recited in claim 3 wherein the switch is a network switch coupling a plurality of sources and a plurality of targets in a network.

8. The method as recited in claim 1 wherein the merged acknowledgment is formed by logically combining the respective acknowledgments.

9. The method as recited in claim 1 wherein the merged acknowledgment encodes the respective acknowledges to indicate to the source which targets successfully received the multicast information.

10. The method as recited in claim 1 wherein the merged acknowledgment indicates whether all of the targets successfully received the multicast information, the merged acknowledgment not identifying which of the targets successfully received or failed to successfully receive the multicast information.

11. The method as recited in claim 10 wherein the merged acknowledgment includes a single bit indicating whether all of the targets successfully received the multicast information.

12. A networked system comprising:

a sending node;

a plurality of receiving nodes coupled to simultaneously receive multicast information sent from the sending node during a multicast operation and coupled to provide acknowledgments indicating whether the multicast information was successfully received; and

a switching medium coupled to supply the multicast information to the respective receiving nodes simultaneously and to receive and combine the respective acknowledgments into a combined acknowledgment supplied to the sending node.

13. The networked system as recited in claim 12 wherein the networked system includes a switched data network and the switching medium is a network switch.

14. The networked system as recited in claim 12 wherein each acknowledgment comprises a plurality of bits, each bit corresponding to a different node, one bit being set to indicate that a node corresponding to the one bit successfully received the multicast information.

15. The networked system as recited in claim 14 wherein the combined acknowledgment includes a plurality of bits corresponding to multicast targets, each bit of the combined acknowledgment that is set corresponding to a node that successfully received the multicast information.

16. The networked system as recited in claim 12 wherein each acknowledgment comprises a plurality of bits, each bit corresponding to one of a plurality of types of errors.

17. The networked system as recited in claim 16 wherein corresponding bits from respective ones of the acknowledgments are combined in the combined acknowledgment, a bit being set to a first predetermined value in the combined acknowledgment to indicate that one or more of the targets had a particular one of the errors and the bit being set to a second value to indicate that none of the receiving nodes had the particular one of the errors.

18. The networked system as recited in claim 12 wherein the acknowledgments from the plurality of target nodes are provided to the switching medium at a fixed time relative to the sending of the multicast information.

19. The networked system as recited in claim 18 wherein the combined acknowledgment is provided to the source node at a fixed time relative to the sending of the multicast information.

20. The networked system as recited in claim 12 wherein the networked system is pipelined.

21. The networked system as recited in claim 12 wherein the switching medium combines the acknowledgments in response to information in each acknowledgment packet that indicates a multicast acknowledge is being sent.

22. The networked system as recited in claim 12 wherein the switching medium combines the acknowledgments into the combined acknowledgment if the acknowledgments arrive at the same time in the switching medium and are destined for a same source.

23. The networked system as recited in claim 12 wherein the switching medium combines the acknowledgments in response to having scheduled a multicast data transfer.

24. The networked system as recited in claim 12 wherein the networked system is operable to reserve switch paths for forwarding the acknowledgments based on switch settings used for forwarding the multicast information.

25. The networked system as recited in claim 12 wherein the networked system includes a plurality of hosts, each of the hosts including both a sending node and a receiving node coupled to the switching medium.

26. An apparatus for transmitting information between an initiator node and a plurality of target nodes, comprising:

means for multicasting information to a plurality of the target nodes from the initiator node; and

means for combining received acknowledgments indicating whether the multicast information was successfully received, into a combined acknowledgment and returning the combined acknowledgment to the initiator node.

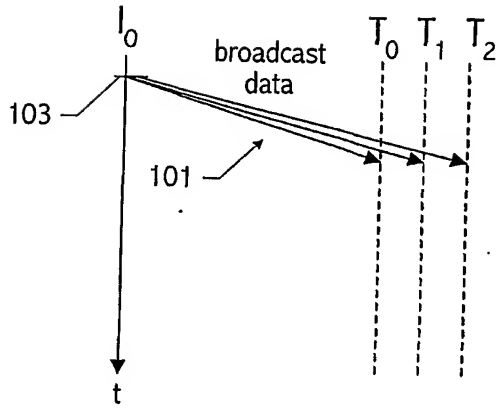

I_0
broadcast
data
T_0 T_1 T_2
103
101
t

FIG. 1

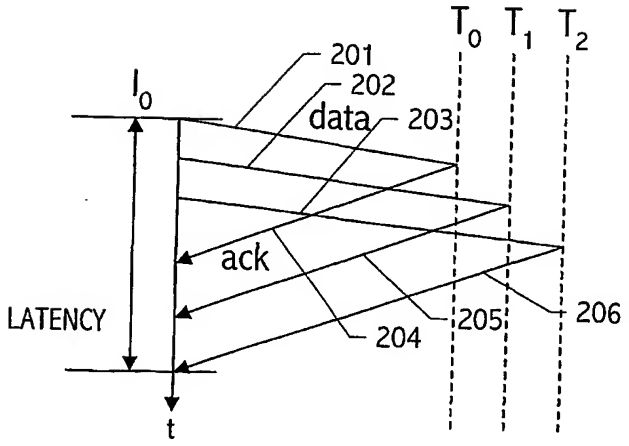

T_0 T_1 T_2
201
202
I_0
data
203
ack
LATENCY
204
205
206
t

FIG. 2

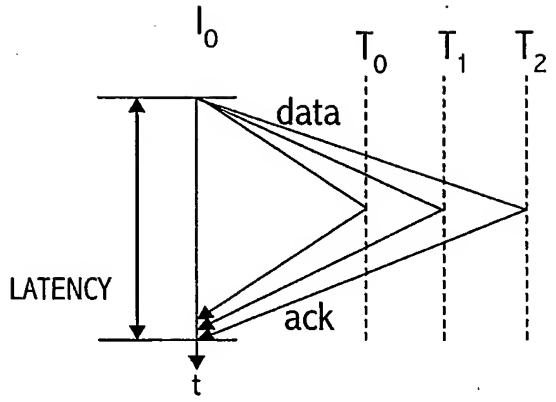

I_0
T_0
T_1
T_2
data
LATENCY
ack
t

FIG. 3

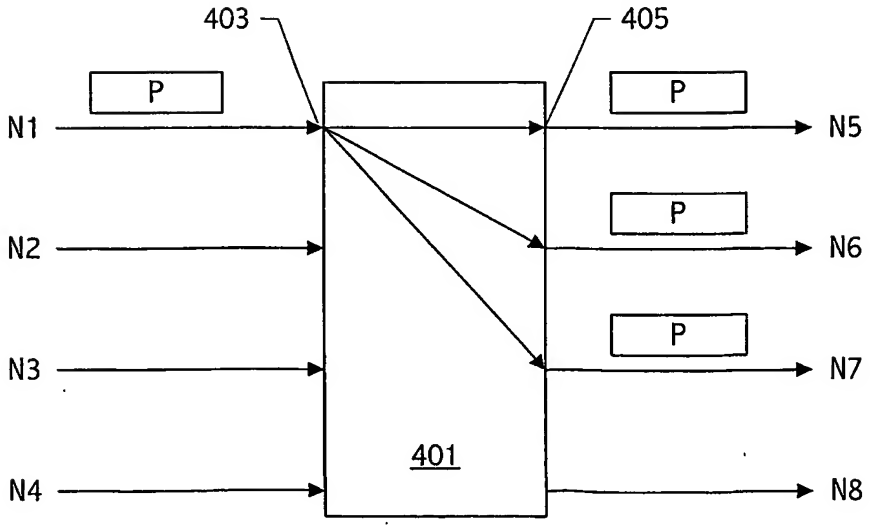

403
405
P
P
N1
N5
P
N2
N6
P
N3
N7
401
N4
N8

FIG. 4

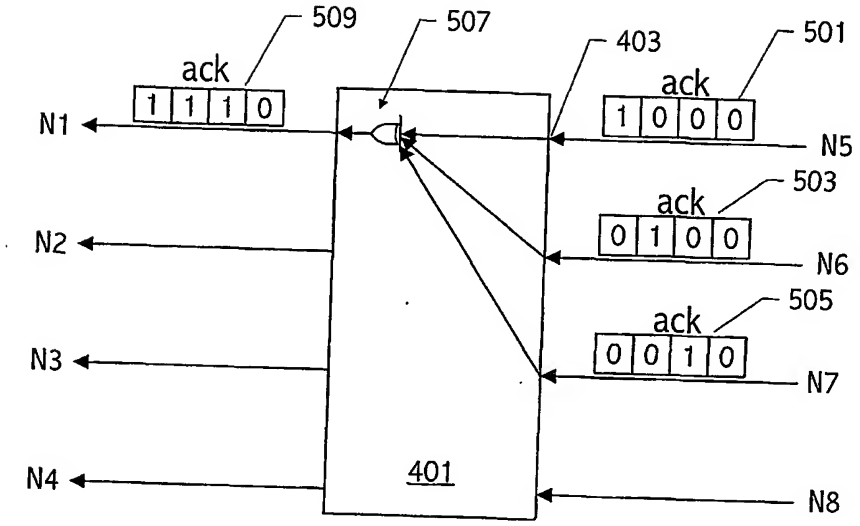

509
507
403
501
ack
1 1 1 0
ack
1 0 0 0
N1
N5
503
ack
0 1 0 0
N2
N6
505
ack
0 0 1 0
N3
N7
N4
401
N8

FIG. 5

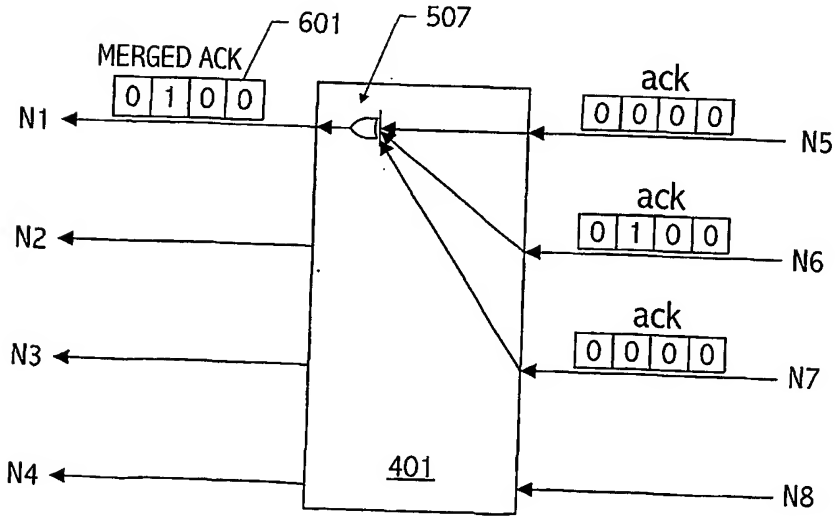

601
507
MERGED ACK
0 1 0 0
ack
0 0 0 0
N1
N5
ack
0 1 0 0
N2
N6
ack
0 0 0 0
N3
N7
N4
401
N8

FIG. 6

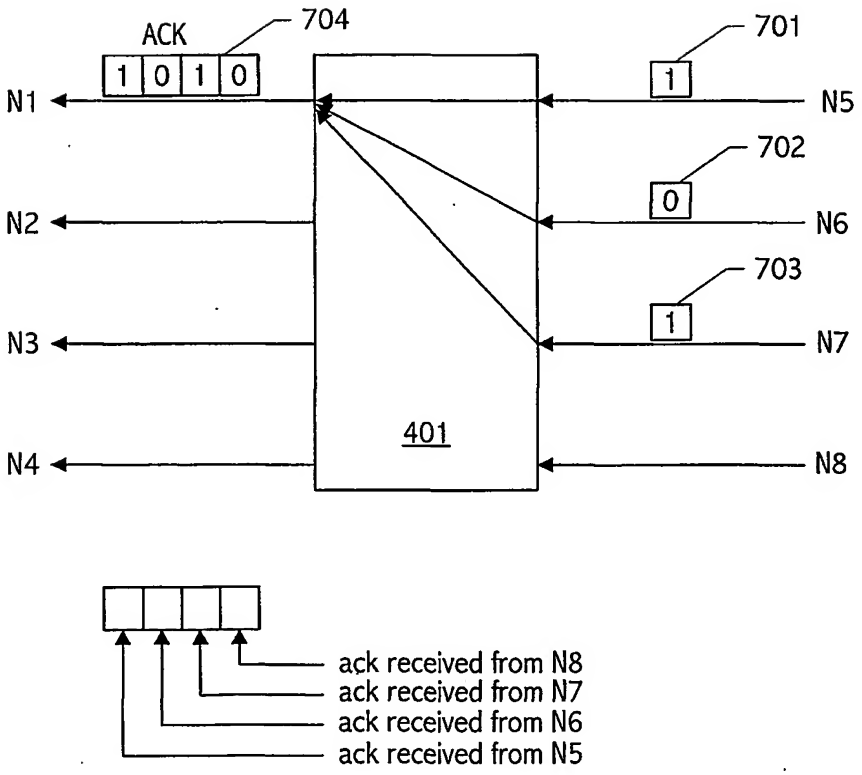

ACK
704
1 0 1 0
701
1
N1
N5
702
0
N2
N6
703
1
N3
N7
401
N4
N8
ack received from N8
ack received from N7
ack received from N6
ack received from N5

FIG. 7

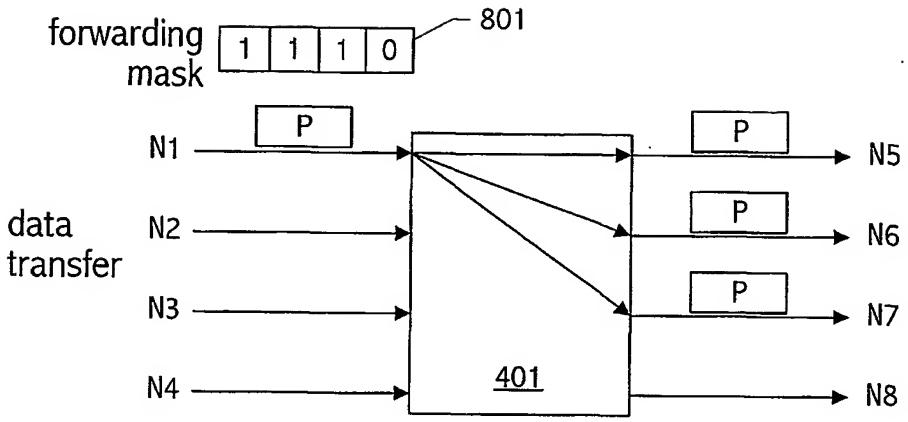

forwarding mask
1
1
1
0
801
P
N1
data transfer
N2
N3
N4
401
P
N5
P
N6
P
N7
N8

FIG. 8

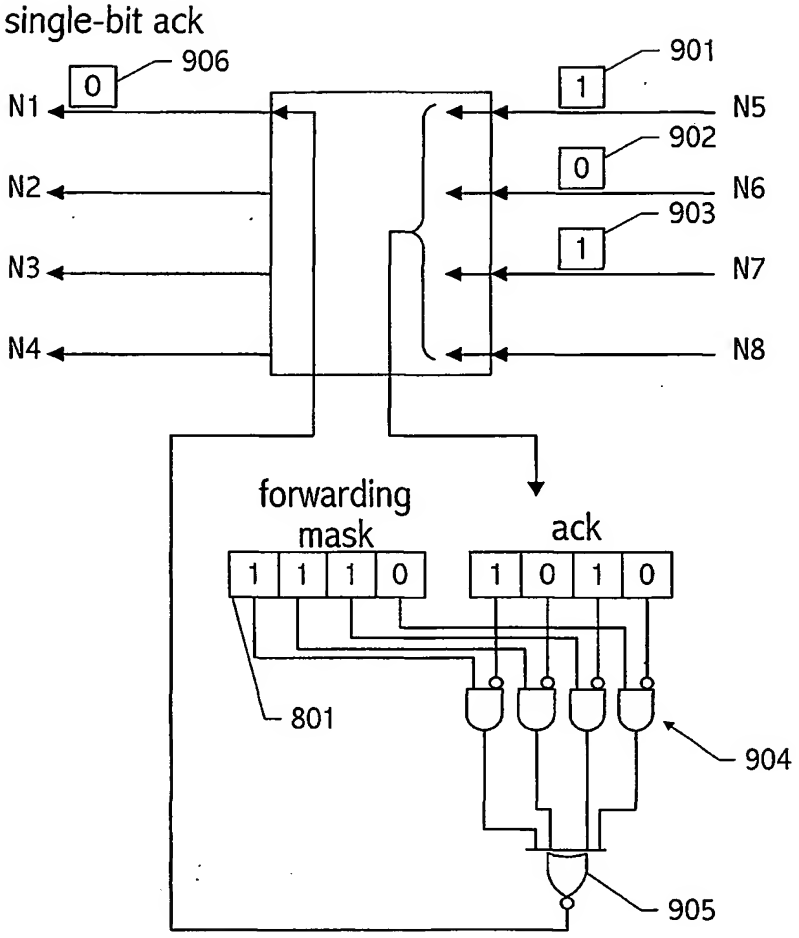

single-bit ack
906
0
N1
N2
N3
N4
901
1
N5
902
0
N6
903
1
N7
N8
forwarding mask
1 1 1 0
ack
1 0 1 0
801
904
905

FIG. 9